Title: Bioinformatic analysis of plant-like vacuole (PLV) proteins in *Toxoplasma gondii*

Author: Javier Tovar

Faculty Sponsor: Paul Ulrich, Georgia State University, Biology Department; Silvia Moreno, University of Georgia, Cell Biology Department

Introduction

The plant-like vacuole (PLV) in *Toxoplasma gondii* is a recently discovered organelle, but little is known about its composition and function. In order to determine the cellular roles of this organelle, fractions enriched in the PLV were analyzed by mass spectrometry (MS) to establish a PLV proteome. Proteomic datasets must be curated to account for the limited resolution of subcellular fractionation and identify contaminants from other cellular compartments. To manage data from the PLV proteome, a database was developed with MySQL to serve as a flexible storage hub through which bioinformatic tools can interact.

Method

To identify proteins resident in the PLV, fractions enriched for the PLV were extracted from wild-type and TgVP1 (a PLV marker) over-expressing strains (Silvia Moreno, UGA).The process divided *T. gondii* lysates into Pellet P3 (complete cells), Fraction 1 (PLV-enriched), and Fraction 6 (non-PLV, cell components). Mass spectrometry was performed on each by collaborators at the Scripps Research Institute. Data were imported to a MySQL database through a series of Perl scripts. To identify contaminants from non-PLV compartments and identify membrane proteins from the PLV, Perl scripts equipped with the Perl-DBI pipelined protein sequences through localization and transmembrane prediction algorithms. A list of putative PLV proteins was derived from an index comparing MS spectral count in Fraction 1 to Fraction 6 and predictions made by subcellular localization algorithms. Proteins were defined as non-PLV contaminants of Fraction 1 based on consensus among two or more algorithms (threshold confidence of ≥80%).

Results

Mass spectrometry analysis of the subcellular fractionation from the TgVP1 Fraction 1 yielded 969 protein groups (total groups in all fractions = 2439). Subcellular predictions identified 179 proteins (18.5%) as likely contaminants from non-PLV compartments. Of the remaining 790 proteins (81.5%), 256 have predicted transmembrane domains.

Discussion

Computational analysis is essential to sort through information produced by proteomic analysis and a critical element of focusing laboratory efforts on high priority research targets. Of the 2448 proteins

identified by MS, 790 are potentially components of the PLV. Roughly 32% of these are putative transmembrane proteins. These analyses are guiding work of collaborators at UGA who are validating localization *in vivo*. The organization and flexibility of this database guarantees ease of use for future analysis, since the database is accessible by most computer languages.